

# Caracterização lexical e sintática de notícias falsas em português produzidas por humanos e por máquinas

Pedro Lucas Castro de Andrade, Renato Moraes Silva, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Inteligência Computacional (NILC)

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

São Carlos – SP – Brasil

pedroandrade@usp.br, {renatoms, taspardo}@icmc.usp.br

## Resumo

Notícias falsas são um grande problema para a sociedade. Com a Inteligência Artificial generativa, notícias falsas produzidas pela máquina têm se proliferado, tornando o cenário mais desafiador. Apesar da relevância desse problema, em línguas sub-representadas como o Português, as pesquisas que buscam diferenciar notícias falsas de humanos e de máquinas são incipientes. Buscando preencher essa lacuna, este artigo explora os corpora Fake.br e FakeTrueBR expandidos com notícias falsas geradas automaticamente, caracterizando lexical e sintaticamente as notícias falsas produzidas por humanos e por máquina. Os resultados mostram que textos gerados por máquina apresentam palavras significativamente mais longas, maior uso de modificadores adjetivais e menor diversidade sintática, apesar de utilizarem mais regras sintáticas por sentença. Em contrapartida, textos humanos exibem maior variabilidade estilística em todas as dimensões analisadas.

## 1 Introdução

Notícias falsas são um grande mal na sociedade moderna, especialmente em função de sua facilidade de disseminação pelas redes sociais. Mais recentemente, com a popularização da Inteligência Artificial generativa, o desafio se tornou ainda maior, dada a facilidade de produção desse tipo de conteúdo por máquinas, com estilo de escrita similar ao dos humanos (Su et al., 2024a; Chen e Shu, 2024).

Silva et al. (2024) compilam diversas iniciativas de detecção automática de notícias falsas para o português, sendo que o Fake.Br (Monteiro et al., 2018; Silva et al., 2020) foi o primeiro corpus de notícias falsas criado para esta língua e o primeiro a ser testado em métodos de detecção. Desde então, diversos esforços surgiram no enfrentamento do problema das notícias falsas.

Apesar dos avanços, pelo que se tem conhecimento, não há iniciativas para o português de tentar diferenciar automaticamente notícias falsas produzidas por humanos das produzidas por máquina. De fato, apenas recentemente o trabalho de Silva et al. (2025) disponibilizou corpora de notícias falsas produzidas por máquina para o português. Para criá-las, os autores utilizaram como base os corpora Fake.br (Monteiro et al., 2018; Silva et al., 2020) e FakeTrueBR (Chavarro et al., 2023), que contêm pares de notícias alinhadas, isto é, para cada notícia falsa escrita por humano, há uma notícia verdadeira relacionada, coletada de agências jornalísticas oficiais. O estudo conduzido por Silva et al. (2025) forneceu essas notícias verdadeiras como entrada para o Sabiá-3 (Pires et al., 2023; Abonizio et al., 2024), um grande modelo de língua (LLM, do inglês *Large Language Model*), solicitando, via *prompt*, a produção de versões falsas. No total, geraram 3.600 notícias falsas para o Fake.br e 1.791 para o FakeTrueBR, igualando o número de notícias falsas produzidas por humanos nesses corpora.

Neste artigo, realiza-se a análise dessas notícias, caracterizando lexical e sintaticamente as notícias falsas produzidas por humanos e por máquina, buscando atributos discriminativos desses textos. Em particular: levanta-se o vocabulário utilizado nessas notícias, avaliando se há diferenças nos termos mais utilizados; analisa-se a complexidade silábica dos textos, comparando o tamanho médio das palavras e sentenças; verifica-se a distribuição de classes gramaticais; extraem-se e se comparam regras sintáticas de formação de sentenças.

Esse esforço é especialmente motivado pela necessidade de identificar a origem do conteúdo falso em uma época em que a Inteligência Artificial generativa tem dominado as discussões científicas e na sociedade em geral. Compreender as diferenças linguísticas entre textos produzidos por humanos e por máquinas é fundamental para o desenvolvimento de sistemas de detecção mais robustos.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a caracterização das notícias falsas; na Seção 4, discutem-se os achados e apresentam-se algumas considerações finais.

## 2 Trabalhos Relacionados

A detecção de notícias falsas e a identificação de textos gerados por LLMs constituem áreas de pesquisa que, embora distintas, têm se aproximado significativamente nos últimos anos. Esta seção apresenta brevemente os principais trabalhos nessas frentes que fundamentam esta pesquisa.

As abordagens tradicionais para detecção de notícias falsas concentram-se em textos humanos e exploram diferentes aspectos linguísticos e contextuais. [Rashkin et al. \(2017\)](#) e [Pérez-Rosas et al. \(2018\)](#) investigaram padrões linguísticos e estilos de escrita como indicadores de falsidade. [Horne e Adalı \(2017\)](#) identificaram que notícias falsas tendem a utilizar vocabulário mais emocional, textos mais simples e tom sensacionalista. Outras abordagens incluem análise semântica e estrutural do conteúdo ([Jin et al., 2017](#); [Zhou e Zafarani, 2020](#)), verificação automática de fatos ([Graves e Cherubini, 2016](#)), análise de credibilidade da fonte ([Baly et al., 2020](#)) e uso de contexto social de redes ([Barnabò et al., 2022](#)). No contexto da língua portuguesa, [Silva et al. \(2024\)](#) apresentam uma visão abrangente dos métodos de detecção, incluindo abordagens baseadas em aprendizado de máquina com atributos linguísticos, modelos de linguagem pré-treinados e técnicas de verificação de conteúdo.

Os métodos de detecção de textos produzidos por máquinas (independentemente da veracidade dos fatos) podem ser categorizados em diferentes abordagens. Métodos baseados em características linguísticas exploram diferenças estatísticas entre textos humanos e gerados por máquina. [Shah et al. \(2023\)](#) demonstraram que características estilísticas como contagem de sílabas, comprimento de palavras e estrutura de sentenças permitem distinguir textos com precisão de 93%. Classificadores baseados em características também têm sido amplamente utilizados, alcançando resultados expressivos por meio da extração de atributos textuais combinados com técnicas de aprendizado de máquina ([Aich et al., 2022](#)).

Métodos baseados em modelos neurais, especialmente *transformers*, têm apresentado os melhores resultados em tarefas de detecção. Modelos como

BERT ([Devlin et al., 2019](#)) e suas variantes são frequentemente usados para classificação, alcançando taxas de precisão superiores a 95% em cenários intra-domínio e superando métodos *zero-shot* como DetectGPT ([Mitchell et al., 2023](#)), apesar de degradarem em domínios não vistos durante o treinamento ([Su et al., 2024b](#)). Deve-se destacar a existência de competições como a AuTextTification, que, em 2024, incluiu o português, sendo que o sistema vencedor utilizou um *ensemble* de *transformers* multilíngues (DistilBERT, mDeBERTa-v3, XLM-RoBERTa) com regressão logística, alcançando Macro-F1 de 0.805 ([Fernández García e Segura-Bedmar, 2024](#)).

Na interseção dessas linhas de pesquisa, [Su et al. \(2023\)](#) demonstraram que detectores de notícias falsas apresentam viés contra textos gerados por LLMs, o que motiva investigações mais profundas. [Su et al. \(2024b\)](#) desvinculam a veracidade da origem do texto, revelando que detectores treinados exclusivamente com textos humanos generalizam melhor para textos gerados por máquina do que o inverso. [Zhou et al. \(2023\)](#) identificaram que textos de IA em redes sociais tendem a ser menos autênticos, mais emocionais e menos analíticos, enquanto textos noticiosos de IA são mais formais e analíticos. Além disso, a IA utiliza menos expressões informais e gírias, mas emprega mais palavras emocionais e expressões de raciocínio. [Berber Sardinha \(2024\)](#) avaliou múltiplos domínios utilizando análise multidimensional e observou que, para notícias, há diferenças sistemáticas em dimensões linguísticas. Para o português, [Silva et al. \(2025\)](#) geraram notícias falsas sintéticas utilizando o modelo Sabiá-3 ([Abonizio et al., 2024](#); [Pires et al., 2023](#)), um LLM específico para português, e avaliaram o impacto em classificadores de aprendizado de máquina. Seus resultados indicam que a natureza dinâmica das notícias falsas e a escassez de recursos para idiomas além do inglês representam desafios significativos para a área.

Inspirado nesses trabalhos, este artigo foca na caracterização linguística comparativa entre notícias falsas escritas por humanos e geradas por LLM. São empregadas principalmente análises lexicais e sintáticas detalhadas para identificar traços discriminativos dos dois grupos, visando fornecer uma análise linguística aprofundada e inédita do fenômeno para o português, assim como subsidiar métodos mais robustos de detecção de notícias falsas no futuro.

### 3 Caracterização das Notícias Falsas

Este trabalho usa os únicos corpora disponíveis para o português que contam com notícias falsas geradas por humanos e por máquina: o Fake.br (Monteiro et al., 2018; Silva et al., 2020) e o FakeTrueBR (Chavarro et al., 2023), cuja geração de notícias falsas por máquina foi conduzida por Silva et al. (2025).

Em particular, os dados do Fake.br foram pré-processados, fazendo-se adequações de formato, como correção de codificação de símbolos especiais e ajuste na segmentação sentencial e na tokenização, deixando os corpora em igualdade de condições para análise e comparação. Isso foi necessário em função das origens diferentes dos corpora e suas decisões de coleta e armazenamento.

A Tabela 1 apresenta as características básicas dos corpora, já separados pela origem das notícias falsas<sup>1</sup>. Nota-se que há variações importantes: as notícias do corpus FakeTrueBR são menores do que as do Fake.br, e as notícias produzidas por máquina são consideravelmente mais extensas do que as humanas em ambos os corpora (368,78 contra 191,14 *tokens* por notícia no Fake.br). Além disso, as sentenças geradas por máquina tendem a ser mais longas (20,57 contra 15,58 *tokens* por sentença no Fake.br). Essas diferenças são levadas em consideração nas análises feitas a seguir.

O Quadro 1 mostra notícias falsas do corpus Fake.br produzidas por humano e por máquina. O texto humano apresenta linguagem informal e emotiva, com uso de exclamações (“*Vergonha!*”), gírias (“*calote*”) e opiniões explícitas do autor (“*roubarem descaradamente*”). Em contraste, o texto gerado por máquina adota um tom formal, com estruturas sintáticas mais elaboradas, uso de fontes anônimas (“*Segundo fontes exclusivas*”) e adjetivos intensificadores (“*suposta dívida milionária*”). Ademais, o texto de máquina é mais extenso e apresenta maior coesão textual, enquanto o texto humano é mais fragmentado e direto.

#### 3.1 Vocabulário

Conduziu-se um levantamento das palavras mais frequentes nas notícias falsas produzidas por humanos e por máquina. A análise vocabular é motivada pelo resultados de Horne e Adali (2017), que mostraram que textos falsos humanos tendem a utilizar

<sup>1</sup>É importante ressaltar que, devido ao pré-processamento realizado, os valores aqui relatados podem divergir dos reportados por Silva et al. (2025).

vocabulário mais emocional e sensacionalista, enquanto Zhou et al. (2023) observaram que textos noticiosos gerados por IA empregam menos expressões informais e mais palavras de raciocínio. Essas evidências sugerem que diferenças vocabulares podem servir como indicadores da origem humana ou de máquina das notícias falsas.

Para tanto, a fim de mitigar ruídos, todas as palavras foram normalizadas para letra minúscula. Na Tabela 2, são mostradas as dez palavras mais frequentes nos corpora (ordenadas das mais para as menos frequentes). Nota-se que há alta sobreposição entre os conjuntos, naturalmente dominados por preposições, artigos e conjunções (consideradas *stopwords*). Não obstante, destaca-se a presença mais recorrente da palavra “como” nos textos gerados por máquina nos dois corpora. Em um primeiro olhar, essa listagem pode facilmente ser considerada irrelevante para a discriminação da origem da notícia, mas é importante lembrar que métodos clássicos de detecção de autoria consideram que essas palavras ajudam a identificar marcas estilísticas de autores, facilitando sua detecção (Mosteller e Wallace, 1964; Stamatos, 2009), e talvez esse também seja o caso para a origem das notícias falsas.

Também foram levantadas as palavras mais frequentes após a remoção das *stopwords* (utilizando-se o NLTK), como pode ser visto na Tabela 3, buscando-se evidenciar mais claramente os conteúdos temáticos. Nesse caso, não foi possível observar discrepâncias entre as notícias.

Em uma análise mais refinada, fazendo-se um levantamento por classe de palavras, utilizando o etiquetador morfossintático porttagger (Silva et al., 2023), observam-se algumas informações lexicais interessantes:

- no Fake.br, há 49% de sobreposição de substantivos entre as notícias de humanos e máquinas e, no FakeTrueBR, esse valor cai para 35%, sendo que os lemas das palavras em comum que estão entre os mais frequentes nesses corpora são “presidente”, “ano” e “governo”;
- para verbos, esses valores são de 45% e 36%, respectivamente, sendo que os lemas das palavras em comum que estão entre os mais frequentes nesses corpora são “poder”, “ter”, “afirmar”, “dizer”, “fazer”, “revelar” e “receber”;

Tabela 1: Características dos corpora.

Corpus	Núm. de notícias	Tokens/sentença	Tokens/notícia	Taxa type/token
Fake.br - Humano	3.600	15,58	191,14	0,5752
Fake.br - Máquina	3.600	20,57	368,78	0,5580
FakeTrueBR - Humano	1.791	17,19	163,17	0,5601
FakeTrueBR - Máquina	1.791	19,92	277,59	0,5846

Quadro 1: Exemplo de notícia falsa produzida por humano (à esquerda) e por máquina (à direita).

Humano	Máquina
Vergonha! Concessionária corta a energia do Maracanã por falta de pagamento. O consórcio Maracanã/Odebrecht deu um “calote” na Light. Desde setembro do ano de 2016, as contas de energia não são pagas. A dívida total ultrapassa R\$ 3 milhões de reais. Na última copa do mundo, que foi realizada no Brasil, foram gastos aproximadamente R\$ 1,3 bilhão no estádio. Nossos políticos, além de roubarem descaradamente o dinheiro público, agora querem tirar os momentos de lazer dos brasileiros [...] no caso o futebol!	Urgente: Luz do Maracanã cortada devido a suposta dívida milionária! A concessionária responsável pelo fornecimento de energia, Light Rio, anunciou que o corte foi necessário após meses de inadimplência. Segundo fontes exclusivas, a dívida teria ultrapassado R\$ 5 milhões, envolvendo não apenas o estádio, mas também áreas adjacentes que não pagam suas contas desde a Olimpíada de 2016. A situação teria se agravado com a nova administração, que assumiu o controle em novembro e não teria honrado os compromissos financeiros. A prefeitura do Rio de Janeiro, questionada sobre o caso, não se pronunciou até o momento, gerando especulações sobre a gestão e o futuro de eventos no icônico estádio.

Tabela 2: Dez palavras mais frequentes nos corpora.

Fake.br (Humano)	Fake.br (Máquina)	FakeTrueBR (Humano)	FakeTrueBR (Máquina)
a	a	a	a
com	as	as	as
de	com	com	com
do	<b>como</b>	da	<b>como</b>
e	de	de	de
em	e	e	e
foi	em	em	em
na	não	não	não
no	o	o	o
não	os	os	os

Tabela 3: Dez palavras mais frequentes nos corpora (sem *stopwords*).

Fake.br (Humano)	Fake.br (Máquina)	FakeTrueBR (Humano)	FakeTrueBR (Máquina)
acordo	afirmou	agora	agora
ainda	agora	anos	ainda
anos	ainda	bolsonaro	apenas
brasil	anos	brasil	bolsonaro
dilma	apenas	dia	brasil
disse	brasil	gente	enquanto
durante	corrupção	governo	especialistas
federal	decisão	lula	governo
governo	enquanto	não	lula
lula	federal	pessoas	mensagem

- para adjetivos, tem-se respectivamente 43% e 28%, sendo que os lemas das palavras em comum que estão entre os mais frequentes nes-

ses corpora são “novo”, “político”, “grande”, “social”, “brasileiro” e “exclusivo”.

É importante esclarecer que esse tipo de levantamento lexical é certamente afetado pela época coberta pelas notícias dos corpora e pelos tópicos mais presentes na mídia nesses momentos, que não são exatamente iguais, fazendo com que essas informações mais específicas mostradas acima sejam mais interessantes como caracterização dos dados do que como traços discriminativos da origem das notícias. No Fake.br, que contém a informação temporal das notícias (ver Figura 1), verificou-se que aproximadamente 88% das notícias falsas foram publicadas nos anos de 2016 e 2017, enquanto que cerca de 90% das notícias verdadeiras foram publicadas em 2017 e 2018. Isso indica uma concentração em tópicos políticos relacionados ao período da Operação Lava Jato e das eleições presidenciais de 2018. Por outro lado, o estudo de [Chavarro et al. \(2023\)](#) informou que as notícias da FakeTrueBR foram publicadas entre 2017 e 2023. Portanto, o FakeTrueBR abrange um período temporal mais amplo, incluindo não apenas notícias políticas, mas também eventos como a pandemia da COVID-19, o que pode ter contribuído para uma maior diversidade temática. Nesse sentido, abordagens futuras de modelagem de tópicos podem ser interessantes de serem aplicadas nos corpora.

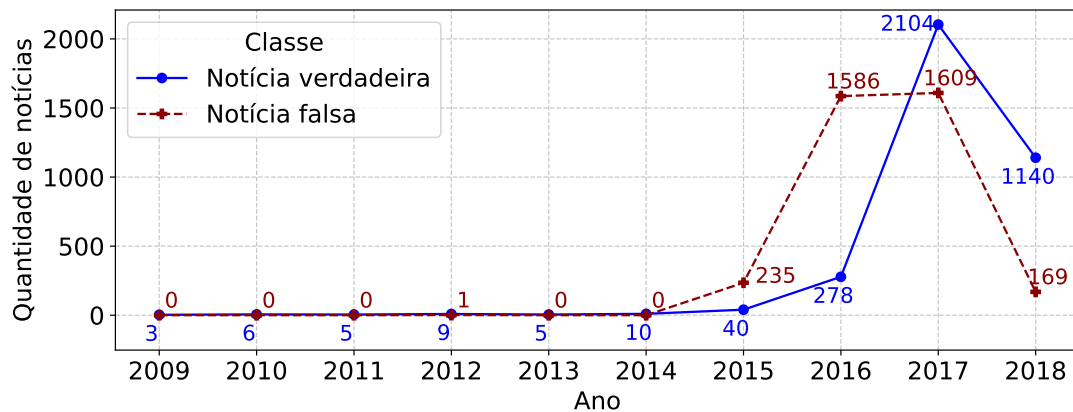


Figura 1: Distribuição temporal das notícias verdadeiras e falsas no corpus Fake.br.

Também foi aplicado o modelo SAGE (*Sparse Additive Generative Model*) proposto por [Eisenshtein et al. \(2011\)](#) para identificar os termos mais distintivos dos textos humanos e de máquina. Diferentemente de abordagens baseadas em frequência absoluta, o SAGE modela a probabilidade de ocorrência de palavras como desvios esparsos em relação a uma distribuição de fundo comum, utilizando regularização L1 para induzir esparsidade nos parâmetros. Essa característica torna o modelo robusto contra ruído estatístico, permitindo que apenas os termos verdadeiramente distintivos apresentem valores significativos. Antes do cálculo, os textos passaram por normalização: conversão para minúsculas e remoção de acentos e cedilha. A distintividade de cada termo é calculada como a diferença entre os desvios específicos de cada classe, de modo que valores positivos indicam termos mais característicos de textos humanos, enquanto valores negativos indicam termos mais frequentes em textos gerados por LLM. A magnitude do valor reflete o grau de distintividade do termo. Na análise feita neste estudo, foram considerados tanto unigramas quanto bigramas para capturar padrões lexicais e colocacionais.

A Tabela 4 apresenta os dez termos mais distintivos para cada grupo. Os resultados revelam padrões morfossintáticos acerca dos textos gerados por máquinas, como a predominância de advérbios e adjetivos intensificadores. Esse contraste sugere que o LLM tende a produzir textos com maior carga de modificadores.

### 3.2 Tamanho de Palavras e de Sentenças

Investigou-se também a complexidade lexical dos textos por meio da contagem de sílabas. A escolha dessa métrica é motivada por trabalhos anteriores

Tabela 4: Termos mais distintivos segundo o SAGE (valores positivos indicam termos típicos de textos humanos; valores negativos, de textos gerados por LLM).

Humano		Máquina	
Termo	SAGE	Termo	SAGE
voce quiser	+7,36	fontes anonimas	-6,54
com gente	+7,32	reviravolta surpreendente	-5,88
org entre	+7,32	anonimas	-5,74
quiser sugerir	+7,31	fontes internas	-5,71
gente pelo	+7,28	fontes nao	-5,56
span	+7,14	revelacao bombastica	-5,43
site facebook	+7,13	enfatizando	-5,31
ao boatos	+7,07	acalorados	-5,29
esse artigo	+6,42	debates acalorados	-5,27
artigo uma	+6,20	cruciais	-5,23

que demonstraram sua relevância na distinção entre textos humanos e gerados por máquina. O trabalho de [Shah et al. \(2023\)](#) mostrou que características como contagem de sílabas e comprimento de palavras permitem distinguir esses textos com precisão de 93%. Além disso, [Horne e Adali \(2017\)](#) identificaram que notícias falsas humanas tendem a utilizar vocabulário mais simples, enquanto [Muñoz-Ortiz et al. \(2024\)](#) observaram que textos gerados por LLMs apresentam distribuições de comprimento de sentença menos dispersas do que textos humanos. Essas evidências sugerem que métricas de complexidade silábica podem capturar diferenças estilísticas discriminativas entre os dois grupos. A Tabela 5 apresenta os resultados agregados dos dois corpora. As diferenças entre os grupos foram avaliadas por meio do teste t de Student, reportando-se o valor da estatística ( $t$ ), o p-valor e o tamanho de efeito medido pelo  $d$  de Cohen.

Nota-se que a máquina produz, em média, palavras significativamente mais longas do que os humanos (2,39 vs 2,17 sílabas por palavra). Essa

Tabela 5: Complexidade silábica nos corpora.

Métrica	Humano	Máquina	<i>t</i>	<i>p</i>	<i>d</i> de Cohen
Sílabas/sentença	44,91 ± 32,97	52,48 ± 9,93	-16,14	< 0,001	0,31 (pequeno)
Sílabas/palavra	2,17 ± 0,14	2,39 ± 0,09	-97,06	< 0,001	1,87 (grande)

diferença apresenta tamanho de efeito grande ( $d = 1,87$ ), indicando que os textos gerados por LLM tendem a utilizar palavras com maior complexidade silábica, o que sugere um vocabulário mais elaborado.

Observa-se que os textos humanos apresentam maior variabilidade na extensão das sentenças (desvio padrão de 32,97 sílabas por sentença, contra 9,93 para a máquina). Esse resultado se alinha com [Muñoz-Ortiz et al. \(2024\)](#), que, ao compararem textos jornalísticos em inglês produzidos por humanos e por LLMs, observaram que textos humanos apresentam distribuições de comprimento de sentença mais dispersas e maior variedade vocabular, sugerindo uma maior padronização estilística nos textos gerados automaticamente.

### 3.3 Classes Gramaticais

Com base na etiquetagem do porttagger ([Silva et al., 2023](#)), analisou-se comparativamente a distribuição de classes gramaticais entre os textos de humanos e de máquina. A motivação para essa análise reside nos resultados obtidos pelo estudo de [Zhou et al. \(2023\)](#), no qual identificou-se que textos noticiosos gerados por IA tendem a ser mais formais e analíticos, com menor uso de expressões informais e maior emprego de palavras emocionais. Essas evidências sugerem que a distribuição de classes gramaticais pode revelar padrões estilísticos distintos entre os dois grupos.

A Figura 2 mostra a proporção de cada classe gramatical nos corpora agregados. A proporção é calculada pelo número de ocorrências de cada classe dividido pelo número total de ocorrências de todas as classes em cada conjunto. É importante avaliar a proporção (e não a frequência absoluta), pois os textos têm tamanhos variados.

Observa-se, por exemplo, que os humanos usam proporcionalmente mais verbos (VERB e AUX), pronomes (PRON) e numerais (NUM), enquanto as máquinas fazem maior uso de substantivos (NOUN), nomes próprios (PROPN) e adjetivos (ADJ). O maior uso de adjetivos pela máquina (que parece ser a diferença mais relevante) está alinhado com os achados de [Zhou et al. \(2023\)](#), segundo

os quais notícias falsas em inglês produzidas por máquina têm maior teor emocional (o que pode ser sinalizado pelos adjetivos).

Para verificar a significância estatística das diferenças, aplicou-se o teste qui-quadrado de independência. Embora o teste indique diferenças significativas na distribuição das classes ( $\chi^2 = 21.579,28$ ;  $p < 0,001$ ), o *V* de Cramer (0,082) revela que a magnitude dessa associação é negligenciável. Isso sugere que, apesar das diferenças, a estrutura morfosintática dos textos produzidos por humanos e por máquinas é bastante similar.

### 3.4 Sintaxe

Para uma análise das diferenças sintáticas entre os textos, extraíram-se regras de formação de sentenças a partir das anotações de dependência no formato *Universal Dependencies* ([de Marneffe et al., 2021](#)), obtidas pelo uso do portparser ([Lopes e Pardo, 2024](#)), um analisador sintático automático para o português brasileiro. A análise sintática é motivada pelo estudo de [Muñoz-Ortiz et al. \(2024\)](#), cujas observações mostraram que textos humanos apresentam maior variedade no uso de tipos de dependência e constituintes sintáticos em comparação com textos gerados por LLMs. Adicionalmente, o trabalho de [Shah et al. \(2023\)](#) demonstrou que a estrutura de sentenças é um atributo relevante para a distinção entre textos humanos e de máquina. Assim, a extração de regras sintáticas permite investigar se essa menor diversidade estrutural também se manifesta no contexto específico de notícias falsas em português.

Cada regra representa um padrão estrutural que combina a classe gramatical de um *token* com as relações de dependência existentes com seus dependentes, em um formato que estende o formalismo discutido em [Junior e Vale \(2025\)](#). Por exemplo, a regra NOUN(DET/det, \*) indica uma estrutura em que um substantivo (NOUN) tem um dependente que o precede na sentença e que é um determinante (DET), associado a ele pela relação de dependência *det*, enquanto a regra \*(VERB) indica que o verbo é a raiz (*root*) da sentença.

As Tabelas 6 e 7 apresentam as estatísticas das

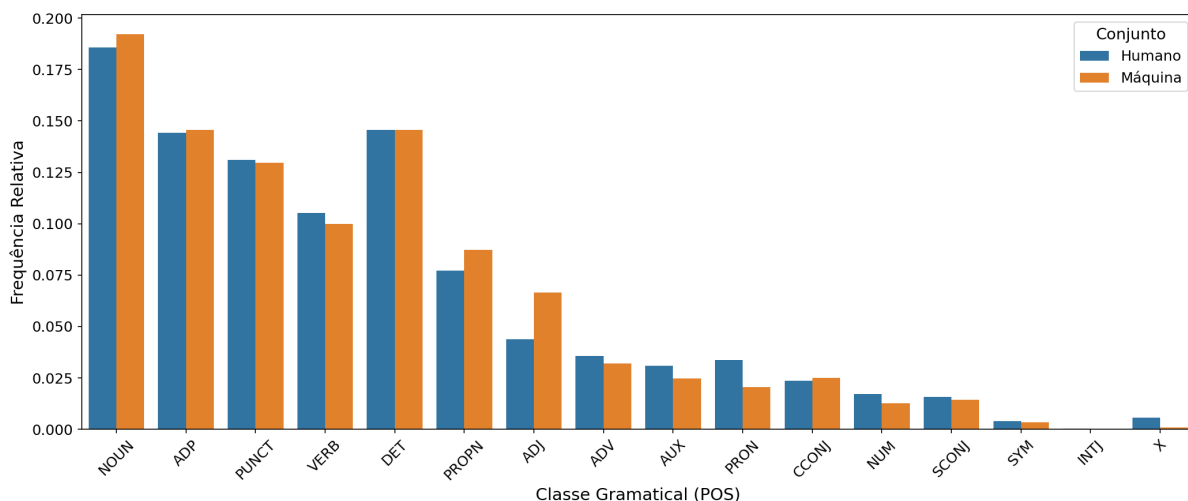


Figura 2: Comparação por classe gramatical nas notícias falsas produzidas por humanos e por máquinas.

regras extraídas. A coluna “Total de regras” indica a quantidade total de regras extraídas, contabilizando as repetições de uma mesma regra dentro de cada sentença. A coluna “Regras únicas” contabiliza apenas as regras distintas, sem repetição. De forma análoga, “Sentenças/texto” é a média de sentenças por notícia e “Regras/texto”, a média de regras por notícia, enquanto “Únicas/texto” indica a média de regras distintas por notícia.

Tabela 6: Regras gramaticais – números absolutos.

Origem	Sentenças	Total de regras	Regras únicas
Humano	50.444	1.595.655	60.177
Máquina	79.170	2.938.295	57.658

Tabela 7: Regras gramaticais – médias.

Origem	Sentenças/texto	Regras/texto	Únicas/texto
Humano	9,36	296,04	74,11
Máquina	14,69	545,04	116,36

Observa-se que, embora a máquina produza mais regras por sentença (37,11 contra 31,63) e mais regras por notícia falsa (545,04 contra 296,04), o número total de regras únicas no corpus é ligeiramente menor (57.658 contra 60.177). Isso sugere maior repetição de padrões sintáticos nos textos gerados automaticamente, indicando uma menor diversidade sintática. Esse achado se alinha com Muñoz-Ortiz et al. (2024), que observou que humanos apresentam maior variedade no uso de tipos de dependência e constituintes sintáticos.

Para verificar a significância estatística dessas diferenças, aplicou-se o teste t de Student, cujos

resultados estão na Tabela 8. Ambas as diferenças são estatisticamente significativas ( $p < 0,001$ ), porém com tamanho de efeito pequeno ( $d = 0,27$  e  $d = 0,38$ , respectivamente). Novamente, nota-se que os textos humanos apresentam maior variabilidade (desvio padrão de 24,62 contra 15,80 para regras por sentença), reforçando o padrão de maior homogeneidade estilística nos textos de máquina.

Para identificar regras mais discriminativas, aplicou-se uma adaptação da técnica TF-IDF (*Term Frequency-Inverse Document Frequency*) sobre as regras extraídas, tratando cada texto como um documento e cada regra como um termo. Foram consideradas apenas regras que possuíssem pelo menos um dependente, excluindo-se regras simples que só introduzissem itens léxicos. Em seguida, comparou-se a distribuição dos *scores* TF-IDF entre os grupos humano e máquina por meio do teste de Mann-Whitney, calculando-se também o  $d$  de Cohen para avaliar o tamanho do efeito. As Tabelas 9 e 10 apresentam as regras mais discriminativas para cada origem, acompanhadas de exemplos de ocorrência. A coluna “Diff” representa a diferença na taxa média de uso da regra entre os grupos (valores positivos indicam regras mais frequentes em textos humanos, enquanto valores negativos indicam regras mais frequentes em textos de máquina).

Observa-se que as regras discriminantes de humanos apresentam tamanhos de efeito predominantemente pequenos ( $d$  entre 0,22 e 0,56), enquanto as regras discriminantes de textos de máquina apresentam efeitos de médios a grandes ( $d$  entre 0,58 e 1,30). Essa assimetria corrobora os achados anteriores sobre a maior homogeneidade estilística dos

Tabela 8: Testes estatísticos para regras de formação de sentenças.

Métrica	Humano	Máquina	<i>t</i>	<i>p</i>	<i>d</i> de Cohen
Regras/sent.	31,63 ± 24,62	37,11 ± 15,80	-48,83	< 0,001	0,27 (pequeno)
Regras únicas/sent.	16,85 ± 7,80	19,41 ± 5,51	-69,21	< 0,001	0,38 (pequeno)

Tabela 9: Regras discriminantes de textos humanos.

Regra	Diff	<i>p</i>	<i>d</i>	Exemplo
PUNCT(PUNCT/punct, *, PUNCT/punct, PUNCT/punct)	+0,012	< 0,001	0,56	<i>No seu governo fez-se o diabo [...] no seu partido</i>
ADV(*, NUM/appos)	+0,006	< 0,001	0,30	<i>O Vaticano iniciou ontem (18) o julgamento de dois ex-dirigentes do hospital infantil da Santa Sé, em Roma.</i>
SYM(ADP/case, DET/det, *)	+0,005	< 0,001	0,29	<i>Além dos R\$ 3,2 milhões para o PT, Bené disse que pegou R\$ 250 mil em dinheiro vivo.</i>
VERB(SCONJ/mark, PRON/nsubj, *, VERB/xcomp, PUNCT/punct)	+0,005	< 0,001	0,29	<i>Se alguém tiver que cair, esse alguém vai ser o Temer.</i>
ADV(ADP/case, *, NUM/appos)	+0,005	< 0,001	0,25	<i>Na manhã de hoje (17) o neto do ex-presidente também cometeu suicídio aos 61 anos.</i>
VERB(*, PUNCT/punct)	+0,005	< 0,001	0,25	<i>Esqueça!</i>
PRON(*, VERB/acl:relcl)	+0,005	< 0,001	0,22	<i>Enfim, veja aos 32:40 Dr. Rey espiando Bolsonaro: Quem dá brechas não pode reclamar que os outros aproveitem, não é mesmo?</i>
VERB(*, NOUN/obj, SYM/obl)	+0,005	< 0,001	0,27	<i>Ele estava com a mulher e queria vender os 2 ingressos por R\$ 2 mil.</i>
NOUN(ADP/case, DET/det, *, NOUN/conj, NOUN/conj)	+0,004	< 0,001	0,26	<i>O garoto foi encaminhado para o hospital com ferimentos nos braços, pernas e rosto.</i>
NOUN(ADP/case, DET/det, *)	+0,004	< 0,001	0,50	<i>Essas pessoas que me expulsaram não servem ao país.</i>

Tabela 10: Regras discriminantes de textos gerados por máquina.

Regra	Diff	<i>p</i>	<i>d</i>	Exemplo
VERB(PUNCT/punct, *, NOUN/obj)	-0,021	< 0,001	1,30	<i>PHA (como é conhecido nas redes sociais) tem uma obsessão doentia: Prender os donos da Rede Globo.</i>
VERB(PUNCT/punct, *, VERB/ccomp)	-0,016	< 0,001	0,79	<i>Esse processo nos levou a explorar a respiração e sua presença dentro de nossas vidas, mostrando que podemos estar sem comida enquanto houver ar.</i>
ADV(ADV/advmood, *)	-0,014	< 0,001	0,71	<i>Fidel não só representa uma Cuba diferente, mas agora parece que ele transcendeu a própria morte!</i>
NOUN(DET/det, *, ADJ/amod, NOUN/nmod)	-0,014	< 0,001	0,68	<i>No Brasil, uma bomba-relógio judicial pode definir o futuro político do país.</i>
PROPON(PUNCT/punct, *, PUNCT/punct)	-0,013	< 0,001	0,66	<i>A decisão também beneficia Frederico Pacheco, primo do senador, e Mendherson Souza Lima, ex-assessor parlamentar de Zeze Perrella (PMDB-MG).</i>
NOUN(*, ADJ/amod)	-0,013	< 0,001	0,73	<i>Definindo Bolsonaro como um "homem autêntico e corajoso", Bivar promete surpresas positivas.</i>
ADV(*, PRON/fixed, PUNCT/punct)	-0,013	< 0,001	0,67	<i>Enquanto isso, a falta de documentação adequada impede famílias como a de Reginaldo de acessar benefícios sociais, como o Bolsa Família.</i>
NOUN(*, ADJ/amod, NOUN/nmod)	-0,013	< 0,001	0,62	<i>Uma megaconferência nos EUA está causando alvoroço ao reunir figuras controversas da política brasileira.</i>
VERB(NOUN/nsubj, *, VERB/ccomp, PUNCT/punct)	-0,012	< 0,001	0,58	<i>Rumores indicam que ele também enfrenta dissidência dentro do chavismo, com Rafael Ramirez emergindo como uma possível alternativa.</i>
NOUN(ADP/case, *, ADJ/amod)	-0,012	< 0,001	0,77	<i>Eles alegam que Bolsonaro é o oposto dos ideais de liberdade econômica e comportamental que desejavam promover.</i>

textos gerados automaticamente.

Entre as regras discriminantes de humanos, destacam-se estruturas de pontuação complexas, como PUNCT(PUNCT/punct, \*, PUNCT/punct, PUNCT/punct), que refletem o uso de reticências ou múltiplos sinais de pontuação típicos de um estilo mais informal. Também se observa maior uso de estruturas com símbolos monetários, como SYM(ADP/case, DET/det, \*), e sentenças curtas imperativas, como VERB(\*, PUNCT/punct).

As regras discriminantes da máquina revelam maior uso de modificadores adjetivais, como NOUN(\*, ADJ/amod) e NOUN(DET/det, \*, ADJ/amod, NOUN/nmod), corroborando a aná-

lise de classes gramaticais. A regra com maior poder discriminativo, VERB(PUNCT/punct, \*, NOUN/obj), apresenta efeito grande ( $d = 1,30$ ), indicando uma forte tendência da máquina em produzir estruturas com objeto direto. A presença de complementos oracionais, evidenciada por VERB(PUNCT/punct, \*, VERB/ccomp) e VERB(NOUN/nsubj, \*, VERB/ccomp, PUNCT/punct), sugere construções sintáticas sofisticadas nos textos de máquina.

#### 4 Discussão e Considerações Finais

As análises realizadas revelam diferenças consistentes entre as notícias falsas produzidas por humanos

e por máquinas, embora com magnitudes variadas. Um padrão recorrente em todas as análises foi a maior variabilidade dos textos humanos. Por exemplo, os desvios padrão para sílabas por sentença (32,97 contra 9,93), regras sintáticas por sentença (24,62 contra 15,80) e regras únicas por sentença (7,80 contra 5,51) foram consistentemente maiores nos textos humanos. Esses resultados reforçam a maior homogeneidade estilística dos textos gerados por máquina. A distribuição de classes gramaticais foi o único caso em que se observou bastante similaridade entre textos de humanos e de máquinas.

Algumas ressalvas e observações são relevantes de serem feitas com relação aos resultados obtidos. Inicialmente, é importante considerar que o método de geração das notícias falsas pela máquina certamente tem impacto nos resultados. Nos corpora utilizados, a máquina utilizou as notícias verdadeiras como base para a geração das falsas. Outros caminhos possíveis incluem a geração sem o uso das notícias verdadeiras como base ou o uso das notícias falsas escritas por humanos como referência. O grande modelo de língua utilizado, o Sabiá-3, também pode influenciar os resultados, sendo sua escolha motivada pelo melhor alinhamento com a língua portuguesa e sua cultura associada. Do ponto de vista de processamento automático, o etiquetador morfossintático e o *parser* de dependências utilizados são do estado da arte para o português brasileiro, mas, como é usual na área, não são perfeitos. Erros pontuais de anotação são esperados, mas não devem alterar os resultados observados. Também é importante destacar que esses sistemas seguem o modelo *Universal Dependencies*, tornando esta investigação mais relevante e, se houver interesse, mais facilmente replicável e comparável com outras línguas.

Outra característica interessante deste estudo refere-se a uma hipótese subjacente, de que discriminar notícias falsas de humanos e de máquinas é diferente da tarefa mais geral de discriminar textos escritos por humanos (não importando se contêm informações verídicas ou não) de textos produzidos por máquina. Conforme discutido anteriormente neste artigo, a literatura aponta nessa direção. Alguns estudos, inclusive, ressaltam que estratégias de escrita diferentes (muitas vezes inconscientes) são utilizadas em notícias falsas, o que sinaliza a relevância de estudos metodológicos diferenciados para esses casos. Além dessa questão, neste artigo não se diferenciaram os casos relacionados de *misinformation*, *disinformation* and *malinformation*,

como definidos por Wardle e Derakhshan (2017), uma vez que os corpora existentes não fazem essa distinção. Porém, esses conceitos consistem em outro ponto relevante que pode interferir na hipótese assumida e influenciar nos resultados.

Por fim, trabalhos futuros incluem a possível continuidade da categorização das notícias, abordando-se os níveis mais desafiadores da semântica e do discurso. Também se pretende avaliar métodos de classificação automática da origem das notícias. Espera-se que a análise apresentada neste artigo possa subsidiar esses métodos e decisões de projeto relacionadas.

## Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Ele também contou com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; processo #2024/17834-6) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; processo #444933/2024-7).

## Referências

- Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, e Ramon Pires. 2024. [Sabiá-3 technical report](#). *Preprint*, arXiv:2410.12049.
- Ankit Aich, Souvik Bhattacharya, e Natalie Parde. 2022. [Demystifying neural fake news via linguistic feature-based interpretation](#). Em *Proceedings of the 29th International Conference on Computational Linguistics*, páginas 6586–6599, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, e Preslav Nakov. 2020. [What was written vs. who read it: News media profiling using text analysis and social media context](#). Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 3364–3374. Association for Computational Linguistics.
- Fabrizio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, e Preslav Nakov. 2022. [Deep learning for cross-lingual news stance detection](#). Em *Proceedings of the ACM Web Conference 2022*, páginas 1094–1103. ACM.
- Tony Berber Sardinha. 2024. [Ai-generated vs human-authored texts: A multidimensional comparison](#). *Applied Corpus Linguistics*, 4(1):100083.

- Juan Pablo Chavarro, Jonata Tyska Carvalho, Tarlis Tortelli Portela, e Jonathan Cardoso Silva. 2023. [Fake-TrueBR: Um corpus brasileiro de notícias falsas](#). Em *Anais da XVIII Escola Regional de Banco de Dados*, páginas 108–117, Porto Alegre. SBC.
- Canyu Chen e Kai Shu. 2024. [Combating misinformation in the age of LLMs: Opportunities and challenges](#). *AI Magazine*, 45(3):354–368.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, e Daniel Zeman. 2021. [Universal dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 4171–4186. Association for Computational Linguistics.
- Jacob Eisenstein, Amr Ahmed, e Eric P. Xing. 2011. [Sparse additive generative models of text](#). Em *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, páginas 1041–1048, Madison, WI, USA. Omnipress.
- Jorge Fernández García e Isabel Segura-Bedmar. 2024. [Human after all: Using transformer based models to identify automatically generated text](#). Em *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, volume 3756 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- Lucas Graves e Federica Cherubini. 2016. [The rise of fact-checking sites in Europe](#). Relatório técnico, Reuters Institute for the Study of Journalism, Oxford, UK.
- Benjamin D. Horne e Sibel Adalı. 2017. [This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news](#). Em *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, páginas 759–766.
- Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, e Jiebo Luo. 2017. [Multimodal fusion with recurrent neural networks for rumor detection on microblogs](#). Em *Proceedings of the 25th ACM International Conference on Multimedia*, páginas 795–816. ACM.
- Isaac Souza de Miranda Junior e Oto Araújo Vale. 2025. [Dependência: o conceito e as gramáticas](#). *Linguamática*, 17(2):71–84.
- Lucelene Lopes e Thiago Alexandre Salgueiro Pardo. 2024. [Towards portparser – a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework](#). Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, páginas 399–409, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, e Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). Em *Proceedings of the 40th International Conference on Machine Learning*, volume 202 de *Proceedings of Machine Learning Research*, páginas 24950–24962. PMLR.
- Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, e Oto A. Vale. 2018. [Contributions to the study of fake news in Portuguese: New corpus and automatic detection results](#). Em *Computational Processing of the Portuguese Language*, volume 11122 de *Lecture Notes in Computer Science*, páginas 324–334, Cham. Springer.
- Frederick Mosteller e David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, MA.
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, e David Vilares. 2024. [Contrasting linguistic patterns in human and LLM-generated news text](#). *Artificial Intelligence Review*, 57(10):265.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, e Rada Mihalcea. 2018. [Automatic detection of fake news](#). Em *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Ramon Pires, Hugo Abonizio, Thales Sales Almeida, e Rodrigo Nogueira. 2023. [Sabiá: Portuguese large language models](#). Em *Intelligent Systems*, volume 14197 de *Lecture Notes in Computer Science*, páginas 226–240, Cham. Springer.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, e Yejin Choi. 2017. [Truth of varying shades: Analyzing language in fake news and political fact-checking](#). Em *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, páginas 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.
- Aditya Shah, Mohand Boughanem, e Gabriella Pasi. 2023. [Feature-based detection of machine-generated text](#). *Expert Systems with Applications*, 228:120321.
- Emanuel Huber Silva, Thiago Alexandre Salgueiro Pardo, e Norton Trevisan Roman. 2023. [Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo “Universal Dependencies”](#). Em *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, páginas 63–73, Belo Horizonte, Brasil. Sociedade Brasileira de Computação.
- Renato M. Silva, Roney L. S. Santos, Tiago A. Almeida, e Thiago A. S. Pardo. 2020. [Towards automatically filtering fake news in Portuguese](#). *Expert Systems with Applications*, 146:113199.

- Renato Moraes Silva, Hazem Amamou, Lucca Baptista Silva Ferraz, Fabio Kauê Araujo da Silva, e Anderson Raymundo Avila. 2025. [Fake news detection in Portuguese under large language model-generated content](#). *Journal of the Brazilian Computer Society*, 31(1):1149–1166.
- Renato Moraes Silva, Roney Lira de Sales Santos, e Thiago Alexandre Salgueiro Pardo. 2024. [Detecção automática de notícias falsas](#). Em Helena de Medeiros Caseli e Maria das Graças Volpe Nunes, editores, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edição, capítulo 27. BPLN.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- Jinyan Su, Claire Cardie, e Preslav Nakov. 2024a. [Adapting fake news detection to the era of large language models](#). Em *Findings of the Association for Computational Linguistics: NAACL 2024*, páginas 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.
- Jinyan Su, Claire Cardie, e Preslav Nakov. 2024b. [Adapting fake news detection to the era of large language models](#). Em *Findings of the Association for Computational Linguistics: NAACL 2024*, páginas 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.
- Jinyan Su, Terry Yue Zhuo, Javid Mansurov, Di Wang, e Preslav Nakov. 2023. [Fake news detectors are biased against texts generated by large language models](#). Em *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, páginas 10276–10288. Association for Computational Linguistics.
- Claire Wardle e Hossein Derakhshan. 2017. *INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making*.
- Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, e Munmun De Choudhury. 2023. [Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions](#). Em *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, páginas 1–20, New York, NY, USA. ACM.
- Xinyi Zhou e Reza Zafarani. 2020. [A survey of fake news: Fundamental theories, detection methods, and opportunities](#). *ACM Computing Surveys*, 53(5):1–40.